

## Information Sciences

## Real-time critical transition discoveries with large language models

Guijun Ma<sup>1,2</sup>, Zidong Wang<sup>3</sup>, Yuzhe Wang<sup>1</sup>, Yong Zhang<sup>4</sup>, Haitao Song<sup>5</sup>, Han Ding<sup>2,6</sup> & Ye Yuan<sup>1,2,\*</sup>

<sup>1</sup>*School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China;*

<sup>2</sup>*State Key Laboratory of Intelligent Manufacturing Equipment and Technology, Huazhong University of Science and Technology, Wuhan 430074, China;*

<sup>3</sup>*Department of Computer Science, Brunel University of London, Uxbridge UB8 3PH, UK;*

<sup>4</sup>*School of Artificial Intelligence and Automation, Wuhan University of Science and Technology, Wuhan 430081, China;*

<sup>5</sup>*Shanghai Artificial Intelligence Research Institute, Shanghai Jiao Tong University, Shanghai 200030, China;*

<sup>6</sup>*School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China*

\*Corresponding author (email: [yue@hust.edu.cn](mailto:yue@hust.edu.cn))

Received 28 September 2025; Revised 8 December 2025; Accepted 10 December 2025; Published online 16 January 2026

**Abstract:** Real-time prediction of critical transitions in complex systems is critical for preventing catastrophic failures and prolonging system lifespans. Existing approaches have extensively focused on qualitative early-warning indicators, but fall short of delivering quantitative, real-time predictions that could guide timely interventions to adjust system states. Here, we present CT-eProber, a large language model-based framework for efficiently probing critical transitions that enables both quantitative and qualitative early warnings. CT-eProber is a general framework that processes prompt data derived from either time-series sensor signals or discrete features, and rapidly adapts to diverse application domains via low-rank adaptation. We demonstrate the effectiveness of the framework on four representative datasets spanning chemistry, finance and robot systems. Results reveal that CT-eProber consistently achieves high predictive accuracy in both real-time quantitative prediction and qualitative classification of critical transitions. Our findings highlight the feasibility of large language model-driven critical transition discovery, establishing a generalizable pathway for real-time prediction and risk prevention in diverse complex systems.

**Keywords:** critical transition, complex system, real-time prediction, large language model

### INTRODUCTION

Early warning of critical transitions in complex systems, ranging from chemistry to finance and robot systems, offers emerging opportunities to prevent unwanted risks or failures [1–4]. For example, in chemistry, real-time prediction of knee points in lithium-ion batteries enables extended cycle lives of batteries by optimizing the usage schedule or charge-discharge protocols [5]; in finance, early detection of systemic financial crises supports proactive policy responses to safeguard economic stability [6]. Over the past few decades, the discoveries of critical transitions have focused on uncovering generic indicators (such as increasing lag-1 autocorrelation, variance, and dynamical eigenvalue) [7,8] or recognizing bifurcation types (such as fold, Hopf, and transcritical bifurcations) [9] in a qualitative manner. Although these approaches

provide valuable theoretical insight, they remain incapable of providing real-time quantitative prediction of when a transition will occur, thereby limiting their practical applicability. A central challenge lies in the fact that system states exhibit minimal observable changes until a critical transition is approached [7,10,11].

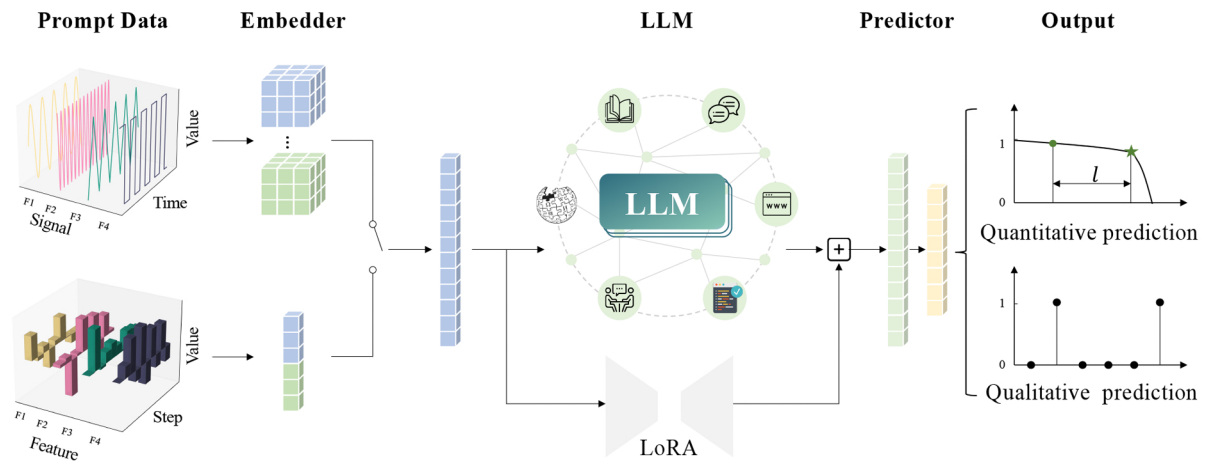
Recent advances in large language models (LLMs) suggest a way forward. Owing to their powerful and generic abilities in logical reasoning, sequence modeling and transfer learning [12,13], LLMs have demonstrated remarkable predictive performance across a wide range of applications, including molecular inverse design [14–16], program synthesis [17], clinical knowledge encoding [18,19], and embodied artificial intelligence [20]. By leveraging transfer learning techniques, pretrained LLMs can be smoothly adapted to downstream tasks using limited domain-specific samples, suggesting their potential for discovering and predicting emergent transitions in complex systems.

Despite this promise, exploiting LLMs for the early prediction of critical transitions faces two major challenges: data heterogeneity and computational burden. First, LLMs are inherently designed to process natural language, whereas complex systems generate diverse data modalities, including time-series sensor signals and discrete features [21–23]. Aligning these modalities requires specialized embedders that transform heterogeneous inputs into unified representations compatible with the LLM backbone [24]. Second, the scale of modern LLMs, often comprising billions of parameters, results in substantial computational demands during fine-tuning and inference [25]. This burden is particularly pronounced in scientific and engineering settings where data availability is limited and computational resources are constrained, highlighting the importance of efficient fine-tuning strategies.

We bridge this gap by introducing CT-eProber (efficient prober of critical transitions), a general framework that leverages LLMs with a resource-efficient fine-tuning strategy to deliver real-time early warning of critical transitions in complex systems. CT-eProber incorporates customized embedders to align heterogeneous input modalities and applies low-rank adaptation (LoRA) to substantially reduce computational costs during downstream fine-tuning. A task-specific predictor further enables either qualitative or quantitative early warning, predicting whether a transition will occur within a specified time horizon or when it will occur. We demonstrate the effectiveness of CT-eProber across three representative domains. In chemistry, it quantitatively predicts knee points of lithium-ion batteries, achieving mean testing errors below 7% across diverse charge-discharge protocols in two representative datasets. In finance, it qualitatively predicts systemic financial crises up to five years in advance, achieving predictive accuracies of 0.97–0.99 and area-under-the-curve (AUC) values as high as 0.96. Moreover, in robot system, CT-eProber provides reliable early warnings of robotic accuracy failure, consistently predicting whether the end-effector error will exceed the critical tolerance within the next four seconds, with accuracies above 0.95 and AUC values of 0.99. These results highlight the feasibility and generalizability of the proposed LLM-driven framework for real-time early warning of critical transitions across diverse complex systems.

## CT-EPROBER

Before presenting applications, we introduce CT-eProber, a framework designed for the early warning of critical transitions in complex systems (Figure 1). CT-eProber comprises four main components: (1) prompt data, (2) embedder, (3) pretrained LLM backbone with resource-efficient transfer, and (4) predictor. Full



**Figure 1** Overview of CT-eProber. Prompt data consist of either observed time-series sensor signals or discrete features collected over a fixed time horizon from complex systems. An embedder aligns heterogeneous inputs, mapping time-series or discrete features into a representation compatible with textual embeddings. The embedded representations are processed by a pretrained LLM, adapted through LoRA for resource-efficient fine-tuning. A task-specific predictor then generates either quantitative or qualitative early warnings of critical transitions. The proposed framework enables real-time prediction of critical transitions across diverse complex systems, which is conducive to mitigating risks and preventing failures.

implementation details are provided in Section “Methods”.

### Prompt data

The design of CT-eProber begins with the formulation of prompt data, which is derived from observed time-series sensor signals or discrete features over a fixed time horizon. For example, the prompt data for lithium-ion battery system consist of four partial charge features [26] from recent cycles (i.e., charge voltage ( $V$ ), charge capacity ( $Q$ ), differential voltage ( $\Delta V$ ) and differential capacity ( $\Delta Q$ )) capturing degradation-related information. For systemic financial crises, the prompt data are constructed from 10-year historical windows of five macro-financial indicators. In the context of robotic accuracy failure, the prompt data consist of six-dimensional joint currents recorded over 100 operating steps (0.8 s). The use of standardized sliding windows ensures comparability across tasks and datasets.

### Embedder

The embedder serves as the interface between the prompt data and the LLM backbone. To bridge the gap between time-series sensor signals or discrete features and the text-trained LLM, CT-eProber incorporates tailored embedders for each data type. For time-series sensor signals, the embedder employs a hierarchical convolutional architecture (Supplementary information Table S1). It consists of three convolution-pooling modules, each composed of a trainable convolutional layer followed by a non-trainable max-pooling layer. This design progressively extracts local temporal features, captures hierarchical patterns, and compresses the information into a compact representation structurally aligned with text tokens. The resulting embeddings preserve degradation trends while filtering out high-frequency noise that could obscure early-warning sig-

nals. For discrete features, the embedder uses a fully connected neural network to project low-dimensional variables into a high-dimensional latent space. This projection enhances the representational richness of the inputs and ensures compatibility with the embedding space of LLM.

## **Pretrained LLM**

The LLM is the core component of CT-eProber, responsible for discovering the inherent dynamical regime with respect to critical transitions of complex systems. We use an LLM named T5 [27], a text-to-text transfer transformer with around three billion parameters, which offers a good balance between computational resources and effectiveness (see Section “Methods”). T5 has been pretrained on a high-quality, diverse, and massive text dataset, and has demonstrated its capabilities on diverse natural language processing tasks such as machine translation, document summarization, and question answering. Importantly, its encoder-decoder architecture and text-to-text training objective showcase powerful sequence-to-sequence modeling capabilities, which have great potential to identify emergent patterns and long-range dependencies, thereby enabling the early warning of critical transitions.

## **Resource-efficient transfer**

Training an LLM, whether starting from scratch or fine-tuning it for downstream tasks, is notorious for its computational burden. To effectively fine-tune T5 for predicting critical transitions under resource constraints, we use a lightweight fine-tuning strategy—LoRA [28], which freezes the pretrained LLM weights and feeds trainable rank decomposition matrices into each LLM layer, thereby substantially reducing the number of trainable parameters. In our experiments, we typically use an NVIDIA GeForce RTX 3090 for training CT-eProber, in which LoRA reduces the number of trainable parameters to just 0.206% of the original size, showcasing its effectiveness in efficiently transferring parameters under resource constraints.

## **Predictor**

The final component of CT-eProber is a task-specific predictor, which translates the latent representations from LLM into predicted values. The predictor design is flexible, enabling both quantitative and qualitative outputs. For quantitative early warning, a regression layer leverages the sequence-to-sequence capacity of T5 to predict future system trajectories. From these trajectories, the predictor estimates the remaining time until a critical transition. For qualitative early warning, the predictor outputs a binary indicator: 0 for a normal condition and 1 for a critical transition. By supporting both quantitative and qualitative tasks within a single architecture, CT-eProber offers a multi-task solution that bridges scientific exploration with practical decision-making.

We now describe some of the new discoveries made by CT-eProber in three different domains, i.e., chemistry, finance, and robot systems. These case studies illustrate its capacity to generalize across scientific domains, leveraging a unified modeling approach while respecting domain-specific characteristics through tailored prompt design.

## RESULTS

We apply CT-eProber to both quantitative and qualitative early warning tasks related to critical transitions. For the quantitative early warning, CT-eProber enables numerical prediction and online detection of the knee point during battery degradation. For the qualitative side, it provides early warnings of systemic financial crises up to five years and of robotic accuracy failures up to four seconds in advance.

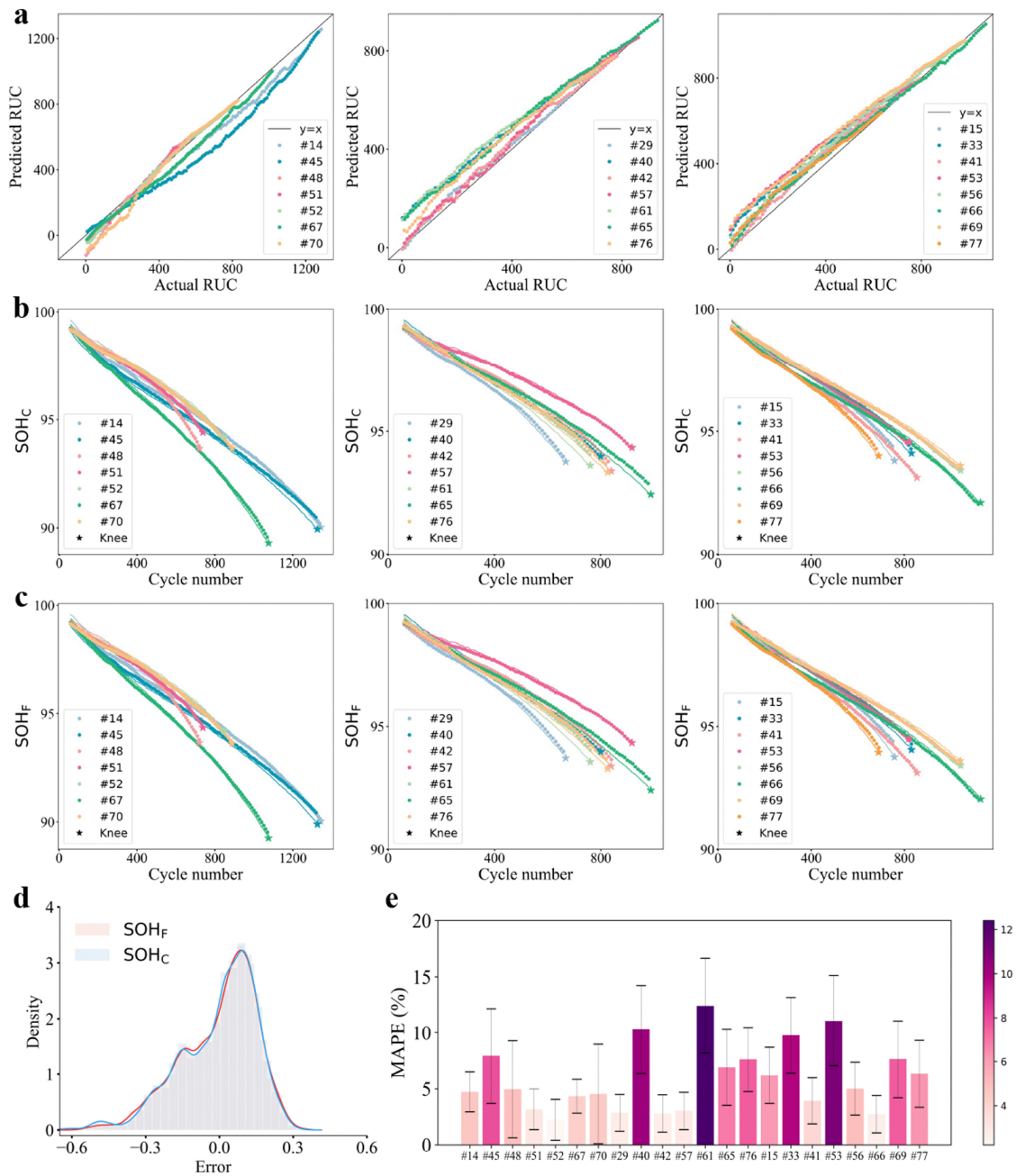
### CT-eProber for quantitative early warning

Lithium-ion batteries are widely recognized as a cornerstone technology for modern energy storage, powering applications from portable electronics to electric vehicles. In recent decades, both academia and industry have devoted increasing attention to monitoring battery operational safety. Particularly, a lithium-ion battery experiences a sudden capacity decline known as the “knee point” (Supplementary information Figure S1), beyond which rapid performance degradation occurs, often accompanied by undesired safety risks. However, in practice, quantitatively predicting the knee point remains challenging due to the inherently complex and nonlinear nature of battery degradation.

Here, we seek to investigate the feasibility of using CT-eProber to real-time predict the knee points in lithium-ion batteries. We validated its quantitative prediction capability on two battery datasets: one generated from our experimental platform (denoted as HUST-LIB) [26] and another publicly released by MIT and Stanford University (denoted as MIT&Stanford-LIB) [29]. HUST-LIB and MIT&Stanford-LIB datasets were generated under distinct discharge and charge protocols, respectively. (1) The HUST-LIB dataset comprises 76 cells exhibiting pronounced knee points, each cycled under a unique multistage discharge protocol. Voltage, current, and capacity were recorded continuously throughout cycling until cell failure. Of these, 54 cells were used to fine-tune CT-eProber, while the remaining 22 cells served as an independent test set for cycle-by-cycle knee-point prediction (Supplementary information Table S2). (2) The MIT&Stanford-LIB dataset, a benchmark for battery health status prediction, comprises 106 cells exhibiting the knee-point phenomenon, cycled under distinct multistage charge protocols. Among the 106 cells, 86 cells were used to fine-tune CT-eProber, and the remaining 20 cells were reserved for independent testing.

During training, CT-eProber is fine-tuned by sliding-window samples, each comprising four partial charge curves ( $V$ ,  $\Delta V$ ,  $Q$ , and  $\Delta Q$ ; see Supplementary information Section 1.1) from the most recent 60 cycles for HUST-LIB and 40 cycles for MIT&Stanford-LIB (see Supplementary information Figure S2). Data from the most recent 10 cycles are fed into the decoder, while earlier cycles are input into the encoder. The primary training label is the number of remaining useful cycles to the knee point (denoted as RUC), determined using the post-event quantile regression (see Section “Methods”) [30]. Two auxiliary targets—the state-of-health at the current cycle (denoted as  $\text{SOH}_C$ ) and over the next 10 cycles (denoted as  $\text{SOH}_F$ )—are incorporated to enhance model tuning as well as predictive performance. Two evaluation metrics, i.e., root mean squared error (RMSE, Eq. (12)) and mean absolute percentage error (MAPE, Eq. (13)), are employed to evaluate the predictive performance of battery knee points.

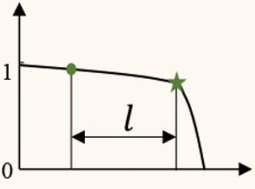
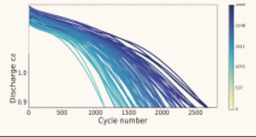
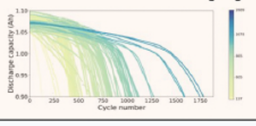



Figure 2a shows the cycle-by-cycle knee point prediction results of CT-eProber on the HUST-LIB dataset, with cells grouped in each subfigure according to the relative error between predicted and actual values. As expected, the predicted RUC values show notably high consistency with the actual values, where CT-e-



**Figure 2** Quantitative early-warning performance of CT-eProber on the HUST-LIB dataset. (a) Cycle-by-cycle prediction of remaining useful cycles (RUC) to the knee point: predicted versus actual values for 22 test cells. Results for intermediate states: (b) State-of-health (SOH) estimation at the current cycle (SOH<sub>C</sub>) versus cycle number, and (c) SOH prediction over next 10 cycles (SOH<sub>F</sub>) versus cycle number. The predicted SOH<sub>F</sub> values are further fed into the post-event quantile regression to detect the knee point in real time, providing a secondary verification step that enhances early-warning reliability. For clarity, results in (a)–(c) are shown at 10-cycle intervals. (d) Relative-error density distributions for SOH<sub>C</sub> and SOH<sub>F</sub>. (e) MAPE of predicted RUC for each test cell.

Prober achieves RMSE of 59 cycles and MAPE of 5.9% on 22 test cells (Table 1). Specifically, Figure 2d exhibits the MAPE with upper and lower bounds for each test cell. In addition, the intermediate model outputs for SOH<sub>C</sub> (Figure 2b) and SOH<sub>F</sub> (Figure 2c) both closely follow the actual SOH trajectories, with

**Table 1** Summary of the quantitative and qualitative early warning results of different datasets

Early warning type	Dataset	Task	Result					
Quantitative early warning 	Battery knee point prediction for HUST-LIB [26] 	<sup>1</sup> RUC	RMSE	59 cycles	MAPE	5.90%		
		<sup>2</sup> SOH <sub>C</sub>		0.14	0.12%			
		<sup>3</sup> SOH <sub>F</sub>		0.14	0.13%			
		<sup>4</sup> KPD		35 cycles	3.40%			
	Battery knee point prediction for MIT&Stanford-LIB [29] 	RUC	48 cycles	7.60%				
		SOH <sub>C</sub>	0.15	0.13%				
SOH <sub>F</sub>		0.15	0.13%					
Qualitative early warning 	Systemic financial crisis prediction [34] 	One-year-ahead	Accuracy	0.97	AUC	0.87	F1 score	0.73
		Two-year-ahead	0.99	0.96	0.80			
		Three-year-ahead	0.97	0.87	0.67			
		Four-year-ahead	0.97	0.84	0.55			
		Five-year-ahead	0.97	0.83	0.62			
	Robotic accuracy failure prediction [36] 	Half-second-ahead	0.97	0.99	0.89			
		One-second-ahead	0.95	0.99	0.92			
		Two-second-ahead	0.99	0.99	0.99			
		Three-second-ahead	0.98	0.99	0.98			
		Four-second-ahead	0.98	0.99	0.99			

Notes: <sup>1</sup>Remaining useful cycles; <sup>2</sup>State-of-health at the current cycle; <sup>3</sup>State-of-health at the next 10 cycles; and <sup>4</sup>knee-point detection

MAPEs of 0.12% and 0.13% for SOH<sub>C</sub> estimation and SOH<sub>F</sub> prediction, respectively (per-cell results are recorded in Supplementary information Table S3), showcasing CT-eProber’s ability to capture nonlinear degradation trends.

Importantly, CT-eProber is not limited to direct prediction of knee points; it also enables online knee-point detection (KPD) in a “prediction-exploration” mode. Specifically, once SOH values for the next 10 cycles are predicted, the quantile regression algorithm will be incorporated to locate the knee point in real time, providing a secondary verification that enhances the early-warning reliability. Across the HUST-LIB dataset, CT-eProber achieves an RMSE of 35 cycles and a MAPE of 3.4% (Table 1) for online KPD, with per-cell results reported in Supplementary information Table S4. It is worth noting that the detected knee points consistently precede the actual values, offering the potential for timely early warning the knee point before the critical transition.

The above results demonstrate the capability of CT-eProber to HUST-LIB dataset with various discharge protocols. Beyond HUST-LIB, CT-eProber can be smoothly adapted to the MIT&Stanford-LIB dataset, which features diverse charge protocols. Using the same fine-tuning strategy (with the encoder window shortened to 30 cycles), CT-eProber achieves the MAPE values of 5.9%, 0.12%, and 0.13% for RUC, SOH<sub>C</sub>,

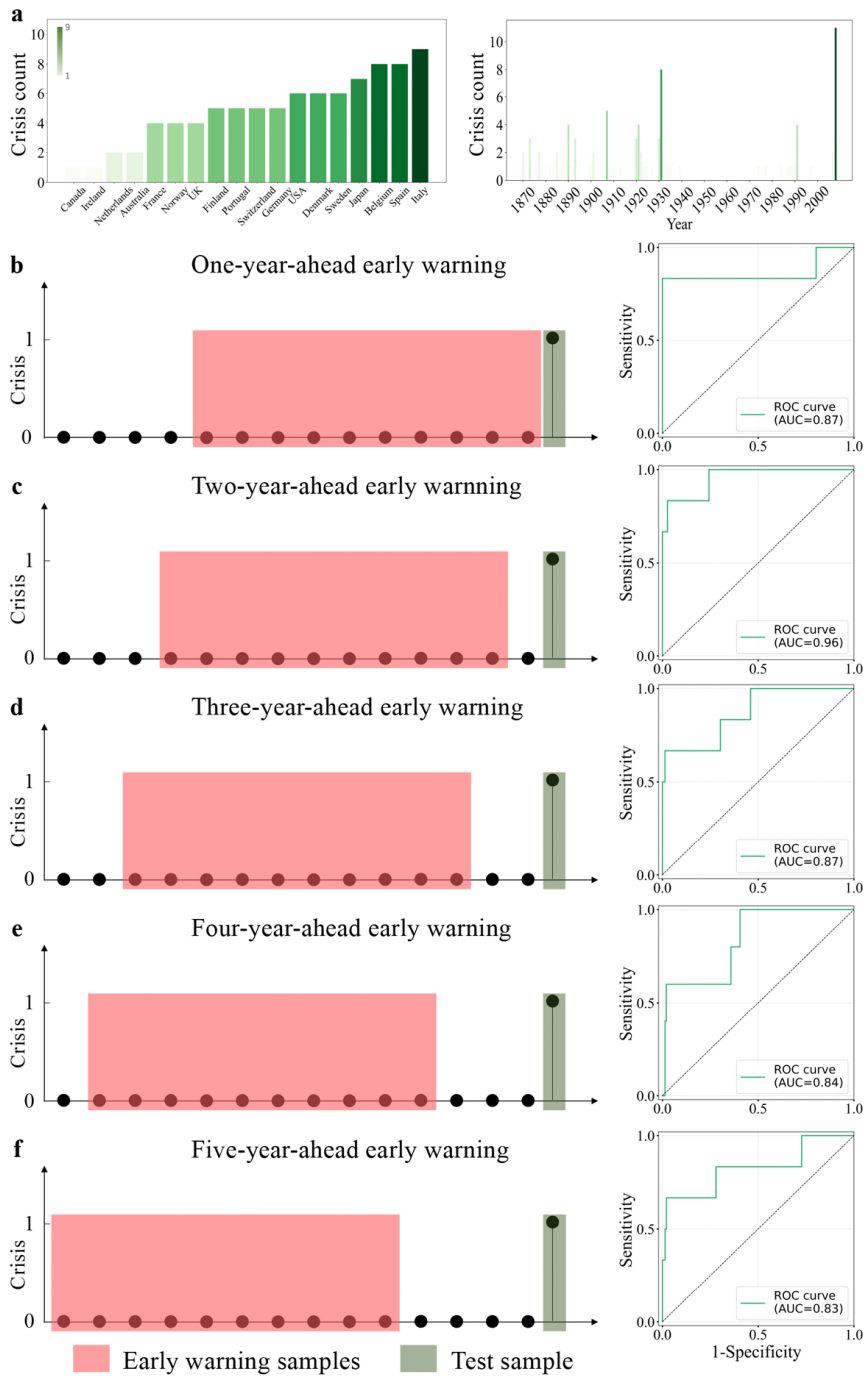
and  $SOH_F$ , respectively, across 20 test cells (Supplementary information Figure S3 and Table S5). Moreover, by leveraging the “prediction-exploration” mode, CT-eProber obtains an overall MAPE of 3.7% for online KPD, with per-cell results provided in Supplementary information Table S6. These findings further underscore the effectiveness and generalizability of CT-eProber for quantitative early warning of the knee point under different battery cycling protocols.

### CT-eProber for qualitative early warning

Systemic financial crises have produced profound and widespread effects, propagating shocks across economic and social domains and imposing heavy burdens on households, firms, and governments [31]. Early warning of such crises is therefore crucial to safeguarding economic stability and mitigating severe social and fiscal costs. A key responsibility for policymakers is the prompt discovery of emerging crises, which enables the deployment of proactive interventions (such as counter-cyclical macroprudential policies) to contain risks before they escalate [32]. To date, early warning of such crises remains a formidable challenge, owing to their complex and multifactorial origins, typically shaped by an interplay of diverse financial factors [6,33].

To this end, we seek to investigate the feasibility of using CT-eProber for the qualitative early warning of systemic financial crises, formulating the task as a binary classification problem in which 0 denotes the normal condition, and 1 represents the occurrence of a crisis. We draw on the systemic financial crisis dataset from the Jordà-Schularick-Taylor Macrohistory Database [34], which spans 18 economies (i.e., Australia, Ireland, Belgium, Canada, Switzerland, Germany, Denmark, Spain, Finland, France, the United Kingdom, Italy, Japan, Netherlands, Norway, Portugal, Sweden, and the United States) over the period 1870–2016. Within this dataset, a total of 88 systemic financial crises and 2630 normal conditions were recorded (Figure 3a). In line with previous study [32], we select five explanatory variables as features: (1) annual growth in loans to the non-financial private sector relative to GDP; (2) annual growth in real stock prices; (3) annual growth in real house prices; (4) the current account-to-GDP ratio; and (5) annual growth in real GDP. Each input sample is generated using a 10-year sliding window, producing a fixed matrix of size  $10 \times 5$ . Features from the most recent six years are fed into the decoder of CT-eProber, while earlier four years are processed by the encoder (see Supplementary information Figure S4). The corresponding label is binary, with 0 denoting a normal condition and 1 indicating the occurrence of a crisis in one specified prediction year. Due to the substantial class imbalance between normal and crisis samples, the dataset is randomly divided into 75% for training and 25% for test using stratified random sampling method [35], and the random over sampling method is applied to alleviate the imbalanced problem during model training. To comprehensively validate the qualitative early warning ability of CT-eProber, we investigate five early-warning tasks, corresponding to one-, two-, three-, four-, and five-year-ahead early warning of systemic financial crises (Figure 3b–f). For instance, the one-year-ahead early warning task indicates that a crisis can be identified one year before its actual occurrence. The performance of CT-eProber is evaluated by its prediction accuracy (Eq. (14)), AUC (Eq. (15)) of the receiver operating characteristic (ROC) and F1 score (Eq. (16)).

Figure 3b–f shows the ROC curves for the five qualitative early-warning tasks of systemic financial crises, with the corresponding AUC values recorded in corresponding legends. Table 1 summarizes the evaluation metrics for these tasks, corresponding to the results shown in Figure 3b–f. The experimental results demonstrate that CT-eProber achieves consistently high performance across all early-warning tasks of systemic financial crises. The model achieves accuracies of 0.97–0.99, underscoring its strong overall early warning



**Figure 3** Qualitative early-warning performance of CT-eProber for systemic financial crises. (a) Distribution of 88 systemic financial crises across countries (left) and years (right) over the period 1870–2016, where the colour denotes the number of crises. (b)–(f) Schematic illustration of one-, two-, three-, four- and five-year-ahead early warning tasks for systemic financial crises, together with the corresponding receiver operating characteristic (ROC) curves.

ability. Among the five tasks, the two-year-ahead early warning exhibits the best performance, with an accuracy of 0.99, an AUC of 0.96, and an F1 score of 0.80, highlighting the model’s capability to balance precision and recall at an optimal horizon. The one- and three-year-ahead tasks also yield robust results, with

AUC values of both 0.87 and F1 scores of 0.73 and 0.67, respectively. In contrast, performance gradually declines for the longer four- and five-year-ahead horizons, where the AUC falls to 0.84 and 0.83, and F1 scores decrease to 0.55 and 0.62, respectively. This trend reflects the increasing difficulty of accurately predicting systemic financial crises over extended forecasting windows, owing to the greater uncertainty and compounding influence of external shocks. Nonetheless, the consistently strong performance across all horizons highlights the potential of CT-eProber as an effective early-warning tool for systemic financial crises.

The results above demonstrate the capability of CT-eProber to provide early warnings of systemic financial crises. Extending beyond the financial domain, we further applied CT-eProber to the real-time early warning of robotic accuracy failure. In this setting, we conducted experiments using a robotic accuracy failure dataset [36], formulating five early warning tasks aimed at determining whether the tool-center-point error would exceed 4 mm (a widely recognized tolerance threshold for end-effector positioning accuracy) within the subsequent  $K$  seconds ( $K = 0.5, 1, 2, 3, 4$ ). These tasks correspond to half-through four-second-ahead early warnings of robotic accuracy failure. Across all horizons, CT-eProber consistently delivers strong performance: prediction accuracies remain above 0.95 (see Supplementary information Figure S5 for confusion matrices), AUC values are close to 1.0 (see Supplementary information Figure S6 for ROC curves), and F1 scores exceed 0.89 (Table 1). Remarkably, predictive performance shows little deterioration as the forecasting window widened, underscoring the robustness and generalization capacity of CT-eProber. Moreover, CT-eProber achieves an inference speed of only 0.0009 s per sample on an NVIDIA GeForce RTX 3090 GPU, fully satisfying the requirements for real-time early warning of robotic accuracy failure. These results further demonstrate the effectiveness of CT-eProber for real-time qualitative early warning across diverse classes of complex systems.

## DISCUSSION

In summary, we have demonstrated the effectiveness and generalizability of CT-eProber for real-time prediction of critical transitions in complex systems. By leveraging a resource-efficient LLM backbone, CT-eProber detects subtle degradation trends and multifactorial risks that conventional approaches often overlook. Its ability to process prompt data constructed from either time-series sensor signals or discrete features enables both qualitative and quantitative early warning across domains. The satisfactory performances achieved in chemistry, finance and robot systems underscore its versatility and highlight broader opportunities for applications in ecosystems, climate, medicine, and other complex systems where real-time prediction of critical transitions is essential.

While CT-eProber has achieved satisfactory performance in this work, there are still some limitations. The framework relies on supervised fine-tuning with labeled critical transition samples, which may be scarce in certain scientific domains. Expanding training with unsupervised, self-supervised or physics-informed strategies may further enhance generalization where labeled samples are limited. Moreover, although CT-eProber achieves satisfactory performance in quantitative early warning, the F1 score of qualitative early warning for systemic financial crisis still leaves room for improvement. In addition, while LoRA sub-

stantially reduces the computational cost, inference with large models still demands considerable resources, raising challenges for deployment on edge devices or embedded systems. Future work should explore smaller specialized foundation models, multimodal pretraining strategies, and uncertainty quantification to improve both efficiency and reliability. Taken together, CT-eProber illustrates the transformative potential of LLMs in critical transition prediction, providing a pathway toward universal, data-driven early-warning systems capable of safeguarding diverse complex systems against catastrophic shifts.

## METHODS

**Datasets.** We validated the performance of CT-eProber in quantitative early warning using two lithium-ion battery datasets (HUST-LIB [26] and MIT&Stanford-LIB [29]), and in qualitative early warning using a systemic financial crisis dataset [34] and a robotic accuracy failure dataset [36]. Both HUST-LIB and MIT&Stanford-LIB consisted of lithium-iron-phosphate (LFP)/graphite A123 APR18650M1A cells (nominal capacity 1.1 Ah, nominal voltage 3.3 V), subjected to different charge-discharge protocols. For quantitative early warning, the HUST-LIB dataset developed in our laboratory comprised 76 cells tested under diverse discharge protocols with a uniform fast-charging protocol, whereas the MIT&Stanford-LIB dataset contained 106 cells cycled under a uniform discharge protocol but subjected to diverse fast-charging protocols. The two datasets provided a comprehensive basis for validating the quantitative early-warning performance of CT-eProber. For qualitative early warning, we employed a systemic financial crisis dataset covering 18 economies from 1870 to 2016, documenting 88 systemic financial crises and 2630 normal conditions, and a robotic accuracy failure dataset comprising 18 complete degradation cycles (six operating conditions, each repeated three times). Detailed information on the four datasets can be found in the Supplementary information Section 1.

**Prompt construction.** For the quantitative early warning of battery knee points, the input prompt consisted of four partial feature curves derived from the most recent  $d$  cycles ( $d = 60$  for HUST-LIB and  $d = 40$  for MIT&Stanford-LIB). Data from the most recent 10 cycles were fed into the decoder of CT-eProber, while earlier cycles were processed by the encoder. The four partial feature curves were extracted within a widely adopted charge range (i.e., 80% SOC to the first 3.6 V) and included: (1) charge voltage curve ( $V$ ); (2) charge capacity curve ( $Q$ ); (3) the differential voltage curve relative to the 10th cycle ( $\Delta V = V_i - V_{10}$ ); and (4) the differential capacity curve relative to the 10th cycle ( $\Delta Q = Q_i - Q_{10}$ ). The four feature curves were parameterized as a function of time and linearly interpolated to a uniform length of 100 to facilitate model manipulations.

For the qualitative early warning of systemic financial crises, the prompt comprised 10 years of historical macro-financial information with five explanatory indicators: (1) annual growth in loans to the non-financial private sector relative to GDP; (2) annual growth in real stock prices; (3) annual growth in real house prices; (4) the current account-to-GDP ratio; and (5) annual growth in real GDP. Features from the most recent six years were provided to the decoder of CT-eProber, while earlier four years were encoded by the encoder. For the qualitative early warning of robotic accuracy failure, the prompt data consisted of six-dimensional joint currents recorded over 100 discrete time steps (0.8 s in total). Of these, the most recent 40 steps were input to

the decoder, whereas the earlier 60 steps were processed by the encoder.

**Model development.** Essentially, CT-eProber takes prompt data  $X$  as input and generates either a numerical prediction or a binary classification  $\hat{Y}$  as output, which is formulated as follows:

$$\hat{Y} = \text{FC}(\text{LLM}(\text{EM}(X))), \quad (1)$$

where the embedder module (EM) is a convolutional neural network for time-series sensor signal prompt data or a fully connected network for discrete feature prompt data, serving to align the heterogeneous sensor measurements with the LLM-dependent textual data. LLM is selected as the pretrained T5, and FC denotes the task-specific regressor or classifier applied on top of the LLM outputs.

For quantitative battery knee-point prediction, the regressor is denoted as

$$\hat{y}_{\text{SOH}_F} = \text{FC}_1(\text{LLM}_d), \quad (2)$$

$$\hat{y}_{\text{SOH}_C} = \text{FC}_3(\text{FC}_2([\text{LLM}_d, \text{LLM}_e])), \quad (3)$$

$$\hat{y}_{\text{RUC}} = \text{FC}_5(\text{FC}_4([\text{LLM}_d, \text{FC}_2([\text{LLM}_d, \text{LLM}_e])])), \quad (4)$$

where  $\text{LLM}_d$  and  $\text{LLM}_e$  represent the decoder and encoder outputs of LLM, respectively, and symbol  $[\cdot, \cdot]$  indicates the concatenation of two vectors.  $\text{FC}_1$ – $\text{FC}_5$  are fully connected layers designed for multiple regression tasks.  $\hat{y}_{\text{RUC}}$  denotes the predicted RUC to battery knee point,  $\hat{y}_{\text{SOH}_C}$  denotes the estimated battery SOH at the current cycle, and  $\hat{y}_{\text{SOH}_F}$  represents the predicted SOH over the next 10 cycles. The overall loss function is defined as a weighted sum of mean squared errors from three regression tasks

$$L_{\text{reg}} = \frac{1}{10 \cdot m} \sum_{i=1}^m \sum_{j=1}^{10} (y_{\text{SOH}_F} - \hat{y}_{\text{SOH}_F})^2 + \frac{\lambda}{m} \sum_{i=1}^m (y_{\text{SOH}_C} - \hat{y}_{\text{SOH}_C})^2 + \frac{\gamma}{m} \sum_{i=1}^m (y_{\text{RUC}} - \hat{y}_{\text{RUC}})^2, \quad (5)$$

where  $m$  represents the number of the training samples, and  $\lambda$  and  $\gamma$  are trade-off parameters introduced to balance the three regression tasks.

For qualitative critical transition classification, the classifier is introduced as

$$\hat{y} = \text{FC}_2(\text{FC}_1(\text{LLM}_d)), \quad (6)$$

where  $\hat{y}$  denotes the predicted probability of a critical transition, with output values close to 0 indicating no transition and values close to 1 indicating the presence of a transition. The classification loss is defined using the binary cross-entropy

$$L_{\text{cls}} = \frac{1}{m} \sum_{i=1}^m (-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)), \quad (7)$$

where  $m$  is the number of training samples,  $y_i \in \{0, 1\}$  is the ground-truth label, and  $\hat{y}_i \in (0, 1)$  is the predicted probability that the  $i$ th sample belongs to a critical transition.

In the model transfer stage, we use a LoRA module to fine-tune CT-eProber for downstream regression and classification tasks (see Supplementary information Figure S10). Specifically, LoRA freezes the pretrained weights of T5 model and injects trainable rank-decomposed matrices into the T5 layers, thereby substantially reducing the number of trainable parameters while preserving model accuracy. For a pretrained weight matrix  $W_0 \in R^{d \times k}$ , LoRA introduces a trainable update  $\Delta W$  that is parameterized through low-rank factorization

$$W = W_0 + \Delta W = W_0 + BA, \quad (8)$$

where  $B \in R^{d \times r}$  and  $A \in R^{r \times k}$ , with rank  $r \ll \min(d, k)$ . During the fine-tuning process, model weights are updated via back-propagation using the AdamW optimizer with a batch size of 32 over 300 training epochs. The learning rate is initialized at 0.001 and reduced by a factor of 0.5 every 20 epochs using a step decay schedule. The model weights are updated until the early-stopping criterion is triggered, i.e., the validation loss shows no improvement for seven consecutive epochs.

**Quantile regression.** In the context of quantitative knee-point early warning for lithium-ion batteries, quantile regression serves both to determine the ground-truth knee point and to enable online detection of the knee point using predicted  $\text{SOH}_F$  values. On the one hand, quantile regression establishes the ground-truth knee points by fitting the historical SOH trajectory and defining a statistically robust safe degradation zone, thereby providing interpretable knee points for CT-eProber training and evaluation (Supplementary information Figure S1). On the other hand, as a secondary verification procedure for early warning of battery knee-point, quantile regression is embedded within the “prediction-exploration” mode of  $\text{SOH}_F$  prediction; once five consecutive residuals of predicted  $\text{SOH}_F$  fall outside the safe zone, the last of these five cycles, together with its corresponding SOH value, is identified as the online-detected battery knee point.

Specifically, a linear model is presented to approximate the SOH trajectory

$$\hat{y}_k = \hat{w}k, \quad (9)$$

where  $k$  denotes the cycle number,  $\hat{y}_k$  is the estimated SOH value, and  $\hat{w}$  is the model coefficient. The coefficient  $\hat{w}$  is obtained by minimizing an asymmetric loss function that penalizes over- and under-estimation differently, which is denoted as

$$\hat{w} = \operatorname{argmin}_w \tau \sum_{y_k > wk} (y_k - wk) + (\tau - 1) \sum_{y_k < wk} (y_k - wk), \quad (10)$$

where  $\tau$  represents the quantile that is strictly between 0 and 1. To establish a statistically robust safe zone, Tukey’s rule [37] is applied to identify and exclude outliers from the residuals,

$$Q_1 - 3 \cdot \text{IQR} \leq \text{res} \leq Q_3 + 3 \cdot \text{IQR}, \quad (11)$$

where  $\text{res}$  is the residual vector, defined as the difference between the SOH values fitted by quantile regression and (i) the ground-truth SOH values for determining the ground-truth knee point, or (ii) the predicted  $\text{SOH}_F$  values for online detection of the knee point.  $Q_1$  and  $Q_3$  are the 25th and 75th percentiles of  $\text{res}$ , respectively, and  $\text{IQR} = Q_3 - Q_1$  is the interquartile range.

**Evaluation.** The performance of CT-eProber in quantitative early warning is evaluated by RMSE and MAPE. For a single cell with  $m$  evaluated cycles, RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{j=1}^m (y_j - \hat{y}_j)^2}, \quad (12)$$

where  $\hat{y}_j$  is the estimated  $\text{SOH}_C$ , predicted  $\text{SOH}_F$  or predicted RUC, and  $y_j$  is the corresponding observed value. MAPE is defined as

$$\text{MAPE} = \frac{1}{m} \sum_{j=1}^m \left| \frac{y_j - \hat{y}_j}{y} \right| \cdot 100\%, \quad (13)$$

where  $y$  denotes the total cycles to the cell knee point for RUC prediction, the observed SOH values for  $\text{SOH}_F$  prediction, or the observed SOH values for  $\text{SOH}_C$  estimation. For multiple test cells, the overall

performance is reported as the average of the RMSE and MAPE across all cells.

The performance of CT-eProber in qualitative early warning is evaluated using the accuracy, AUC, and F1 score, which are defined as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \cdot 100\%, \quad (14)$$

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})d(\text{FPR}), \quad (15)$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (16)$$

where TP and TN denote the number of critical transitions and normal conditions correctly identified, respectively. FP and FN represent the number of false predictions of critical transitions and normal conditions, respectively.  $\text{FPR} = \text{FP} / (\text{TN} + \text{FP})$ ,  $\text{TPR} = \text{Recall} = \text{TP} / (\text{TP} + \text{FN})$  and  $\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$ . The F1 score ranges from 0 to 1, with higher values indicating a better balance between false alarms and missed predictions in early warning of critical transitions.

### Data availability

The original data are available from corresponding authors upon reasonable request.

### Acknowledgements

We are grateful to Mr. Xiaoran Yang and Mr. Xiaoyou Duan for early insightful discussion.

### Funding

This work was supported by the National Natural Science Foundation of China (52188102, T2525012, and 62503185).

### Author contributions

Y.Y. conceptualized the algorithm. G.M. developed the algorithms under the supervision of H.D., Z.W., and Y.Y. G.M. collected and analyzed the lithium-ion battery data and the systemic financial crisis data; Y.W., Y.Z., and H.S. collected and analyzed the robotic accuracy failure data. G.M. and Y.W. wrote the manuscript. All authors designed and discussed the study.

### Conflict of interest

The authors declare that they have no conflict of interest.

### Supplementary information

The supporting information is available online at <https://doi.org/10.1360/nso/20250060>. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

### References

- 1 Lenton TM. Early warning of climate tipping points. *Nat Clim Change* 2011; **1**: 201–209.
- 2 Armstrong McKay DI, Staal A, Abrams JF, *et al.* Exceeding 1.5°C global warming could trigger multiple climate tipping points. *Science* 2022; **377**: eabn7950.
- 3 Veraart AJ, Faassen EJ, Dakos V, *et al.* Recovery rates reflect distance to a tipping point in a living system. *Nature* 2012;

- 481: 357–359.
- 4 Cerini F, Childs DZ, Clements CF. A predictive timeline of wildlife population collapse. *Nat Ecol Evol* 2023; **7**: 320–331.
  - 5 Attia PM, Bills A, Brosa Planella F, *et al.* Review—“Knees” in lithium-ion battery aging trajectories. *J Electrochem Soc* 2022; **169**: 060517.
  - 6 Engle RF, Ruan T. Measuring the probability of a financial crisis. *Proc Natl Acad Sci USA* 2019; **116**: 18341–18346.
  - 7 Scheffer M, Bascompte J, Brock WA, *et al.* Early-warning signals for critical transitions. *Nature* 2009; **461**: 53–59.
  - 8 Grziwotz F, Chang CW, Dakos V, *et al.* Anticipating the occurrence and type of critical transitions. *Sci Adv* 2023; **9**: eabq455.
  - 9 Bury TM, Sujith RI, Pavithran I, *et al.* Deep learning for early warning signals of tipping points. *Proc Natl Acad Sci USA* 2021; **118**: e2106140118.
  - 10 Scheffer M, Carpenter SR, Lenton TM, *et al.* Anticipating critical transitions. *Science* 2012; **338**: 344–348.
  - 11 Kim H, Kim I, Kim M, *et al.* Detection of the knee point in lithium-ion battery degradation using a state-of-charge-dependent parameter. *Proc Natl Acad Sci USA* 2025; **122**: e2424838122.
  - 12 Gruver N, Finzi M, Qiu S, *et al.* Large language models are zero-shot time series forecasters. In: *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*. New Orleans, 2023.
  - 13 Zhou T, Niu P, Wang X, *et al.* One fits all: Power general time series analysis by pretrained LM. In: *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*. New Orleans, 2023.
  - 14 Jablonka KM, Schwaller P, Ortega-Guerrero A, *et al.* Leveraging large language models for predictive chemistry. *Nat Mach Intell* 2024; **6**: 161–169.
  - 15 M. Bran A, Cox S, Schilter O, *et al.* Augmenting large language models with chemistry tools. *Nat Mach Intell* 2024; **6**: 525–535.
  - 16 Madani A, Krause B, Greene ER, *et al.* Large language models generate functional protein sequences across diverse families. *Nat Biotechnol* 2023; **41**: 1099–1106.
  - 17 Romera-Paredes B, Barekatin M, Novikov A, *et al.* Mathematical discoveries from program search with large language models. *Nature* 2024; **625**: 468–475.
  - 18 Singhal K, Azizi S, Tu T, *et al.* Large language models encode clinical knowledge. *Nature* 2023; **620**: 172–180.
  - 19 Thirunavukarasu AJ, Ting DSJ, Elangovan K, *et al.* Large language models in medicine. *Nat Med* 2023; **29**: 1930–1940.
  - 20 Zeng F, Gan W, Wang Y, *et al.* Large language models for robotics: A survey. arXiv: [2311.07226](https://arxiv.org/abs/2311.07226).
  - 21 Jin M, Wang S, Ma L, *et al.* Time-LLM: Time series forecasting by reprogramming large language models. arXiv: [2310.01728](https://arxiv.org/abs/2310.01728).
  - 22 Yuan Y, Ma G, Cheng C, *et al.* A general end-to-end diagnosis framework for manufacturing systems. *Natl Sci Rev* 2020; **7**: 418–429.
  - 23 Yuan Y, Liu J, Jin D, *et al.* DeceFL: A principled fully decentralized federated learning framework. *Natl Sci Open* 2023; **2**: 20220043.
  - 24 Zhang H, Wang Q, Zhang W, *et al.* Estimating comparable distances to tipping points across mutualistic systems by scaled recovery rates. *Nat Ecol Evol* 2022; **6**: 1524–1536.
  - 25 Zheng Z, Ren X, Xue F, *et al.* Response length perception and sequence scheduling: An LLM-empowered LLM inference pipeline. In: *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*. New Orleans, 2023.
  - 26 Ma G, Xu S, Jiang B, *et al.* Real-time personalized health status prediction of lithium-ion batteries using deep transfer learning. *Energy Environ Sci* 2022; **15**: 4083–4094.
  - 27 Raffel C, Shazeer N, Roberts A, *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res* 2020; **21**: 5485–5551.
  - 28 Hu EJ, Shen Y, Wallis P, *et al.* LoRA: Low-rank adaptation of large language models. In: *Proceedings of the 10th International Conference on Learning Representations (ICLR)*. Online, 2022.

- 29 Severson KA, Attia PM, Jin N, *et al.* Data-driven prediction of battery cycle life before capacity degradation. *Nat Energy* 2019; **4**: 383–391.
- 30 Zhang C, Wang Y, Gao Y, *et al.* Accelerated fading recognition for lithium-ion batteries with Nickel-Cobalt-Manganese cathode using quantile regression method. *Appl Energy* 2019; **256**: 113841.
- 31 Gorton G. Financial crises. *Annu Rev Financ Econ* 2018; **10**: 43–58.
- 32 Tölö E. Predicting systemic financial crises with recurrent neural networks. *J Financ Stabil* 2020; **49**: 100746.
- 33 Chen S, Huang Y, Ge L. An early warning system for financial crises: A temporal convolutional network approach. *Tech Econ Dev Eco* 2024; **30**: 688–711.
- 34 Jordà Ò, Schularick M, Taylor AM. Macrofinancial history and the new business cycle facts. *NBER Macroecon Annu* 2017; **31**: 213–263.
- 35 Loss T, Colbrook MJ, Hansen AC. Stratified sampling based compressed sensing for structured signals. *IEEE Trans Signal Process* 2022; **70**: 3530–3539.
- 36 Qiao G, Weiss BA. Accuracy degradation analysis for industrial robot systems. In: *Proceedings of the 12th International Manufacturing Science and Engineering Conference*. Los Angeles, 2017.
- 37 Schwertman NC, Owens MA, Adnan R. A simple more general boxplot method for identifying outliers. *Comput Stat Data Anal* 2004; **47**: 165–174.